arXiv:1302.2023v1 [math.ST] 8 Feb 2013

Random sets and exact confidence regions

Ryan Martin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
rgmartin@uic.edu

February 11, 2013

Abstract

An important problem in statistics is the construction of confidence regions for unknown parameters. In most cases, asymptotic distribution theory is used to construct confidence regions, so any coverage probability claims only hold approximately, for large samples. This paper describes a new approach, using random sets, which allows users to construct exact confidence regions without appeal to asymptotic theory. In particular, if the user-specified random set satisfies a certain validity property, confidence regions obtained by thresholding the induced data-dependent plausibility function are shown to have the desired coverage probability.

Keywords and phrases: Coverage probability; inferential model; plausibility function; predictive random set; validity.

AMS subject classification: 62F25; 60D05; 62E15.

1 Introduction

A fundamental problem in statistics is that of constructing confidence regions. Roughly speaking, a confidence region is a data-dependent subset of the parameter space with the interpretation that, all values inside this subset are "reasonable" estimates of the unknown parameter. The more precise interpretation of confidence regions is based a frequentist notion of coverage probability. That is, in repeated sampling, the confidence region will contain the true parameter value a specified proportion of the time. That the confidence region (nearly) hits the target coverage probability is crucial to the validity of the resulting inference: on one hand, if the actual coverage probability is too high, then the confidence regions are likely too large to provide any meaningful notion of uncertainty; on the other hand, if the actual coverage probability is too low, then it is likely that the confidence region has a systematic bias, casting doubt on the accuracy of the results for the data at hand. Unfortunately, it is rare that a simple and exact confidence region is available; the well-known Student-t confidence interval for a Gaussian mean is one exception. Typically, an appeal to asymptotic theory is made, and confidence regions are built based on the simpler limiting distribution; confidence regions based on the asymptotic normality of maximum likelihood estimators is one example. However, with this approach, one must

add to any coverage probability claim the caveat "for sufficiently large sample size." Alternatively, numerical methods, such as bootstrap (Efron and Tibshirani 1993), are popular when a direct appeal to asymptotic theory is questionable. But validity of the bootstrap also depends on large-sample theory, so there are no non-asymptotic coverage probability guarantees for bootstrap confidence regions.

This paper describes a new approach, resting on a theory of random sets. The initial step is to establish an association between the observable data, the unknown parameter, and a mostly arbitrary auxiliary variable. By "association" here we mean a suitable representation of the statistical model for the observable data. Alternatively, the association can be viewed as a sort of compatibility relation among the various inputs. Random sets supported in the auxiliary variable space—called predictive random sets—are propagated, via observed data and the specified statistical model, to random sets in the parameter space. These random sets in the parameter space are characterized by their belief functions or, alternatively, by their plausibility functions. These functions also appear in the famous Dempster–Shafer theory (Dempster 2008; Shafer 1976), but the approach described here is different. It is shown in Section 4 that, under very mild conditions on the user-specified predictive random sets, exact confidence regions can be constructed via a suitable thresholding of the plausibility function.

The remainder of the paper is organized as follows. Section 2 describes the general statistical problem and defines confidence regions and coverage probability. Random sets are described in Section 3, with a general overview in Section 3.1 and a presentation of the important new concept, namely, predictive random sets, in Section 3.2. These sets are the driving force behind the proposed approach. In Section 4 we first define the plausibility function, which is nothing but a probability calculation relative to the distribution of the predictive random set, along with the corresponding plausibility region. Then we prove the main result that, under mild conditions on the model itself, if the predictive random set is valid, a property that is easily satisfied, then the corresponding plausibility regions hit the desired coverage probability. This is a finite sample result, not asymptotic. Here we find that certain aspects of the formal mathematical theory of random sets leads to a relatively simple statement of the sufficient conditions for this result. Two illustrative examples, involving models used in reliability theory, are presented in Section 5. Finally, Section 6 contains a brief discussion.

2 Setup and notation

Let Y be an observable sample, taking values in the sample space $\mathbb{Y}$, with distribution $\mathsf{P}_{Y|\theta}$ depending on a parameter θ in a parameter space Θ. Here Y may be a vector of n (possibly independent) observations, so that $\mathbb{Y}$ is actually a product space, but it is not necessary to be so specific here. The distribution $\mathsf{P}_{Y|\theta}$ is called the sampling model, and if the value of θ were known, then Y could be simulated. In the present context, the actual θ is unknown, and the goal is to use data Y to make inference about θ.

In statistical applications, it is typical to summarize data Y with a statistic $T = T(Y)$. Just like Y, the statistic T has a sampling distribution, denoted by $\mathsf{P}_{T|\theta}$, which usually depends on θ. In fact, one usually takes T to be a minimal sufficient statistic for θ, though deviation from this guiding principle is sometimes warranted; see Section 5.1. The

initial reduction of X to a minimal sufficient statistic T can be justified by the standard arguments of Fisher or, more generally, by those of Martin and Liu (2012). The classical frequentist approach to statistical inference derives procedures, such as hypothesis tests, based on the sampling distribution of T. In this paper, focus is on confidence regions. Let $\mathscr{C}_\alpha(T)$ be a T-dependent subset of Θ. For given $\alpha \in (0,1)$, $\mathscr{C}_\alpha(T)$ is called a $100(1-\alpha)\%$ confidence region for θ if

$$\mathsf{P}_{T|\theta}\{\mathscr{C}_\alpha(T) \ni \theta\} \geq 1-\alpha, \quad \forall\, \theta \in \Theta. \tag{1}$$

The left-hand side of (1) is the coverage probability of $\mathscr{C}_\alpha(T)$, and the definition of confidence region places a condition on this coverage probability, namely, that it must exceed the $1-\alpha$ level. In words, (1) states that, if the confidence region $\mathscr{C}_\alpha(T)$ is used in many examples involving data $Y \sim \mathsf{P}_{Y|\theta}$ and statistic $T = T(Y)$, then roughly $100(1-\alpha)\%$ of the realized regions will contain the true parameter value. In other words, if α is small, i.e., $\alpha = 0.05$, then $\{\mathscr{C}_\alpha(T) \not\ni \theta\}$ is a rare event with respect to the sampling distribution of T. So, in practice, users will use this "rare event" interpretation to justify the conclusion that their calculated confidence region contains the true θ value.

Clearly, it is most efficient for the $100(1-\alpha)\%$ confidence region to have coverage probability equal $1-\alpha$; this would indicate that, in some sense, its size is just right. In practice, however, for the sake of analytical or computational convenience, this efficiency is sacrificed. That is, confidence regions used in practice may not exactly satisfy (1). Equality may hold in (1) only as $n \to \infty$, and for finite n, the true coverage probability may be above or below the desired $1-\alpha$ level. It would desirable to have a general way to construct regions $\mathscr{C}_\alpha(T)$ that satisfy (1) for all n, especially if equality can be attained in some cases. The objective of this note is to present and justify such a construction.

Towards this, we must first digress a bit to introduce an alternative representation of the sampling model $\mathsf{P}_{T|\theta}$, one that involves an auxiliary variable. Let $\mathbb{U}$ be an (arbitrary) auxiliary variable space, equipped with a probability measure P_U. Then choose a function $a : \mathbb{U} \times \Theta \to \Theta$, such that, if $U \sim \mathsf{P}_U$, then $a(U,\theta) \sim \mathsf{P}_{T|\theta}$. In other words, the sampling distribution of T can be characterized by the following recipe:

$$\text{sample } U \sim \mathsf{P}_U \text{ and set } T = a(U,\theta). \tag{2}$$

This is a familiar notion in the context of simulation, e.g., the inverse probability transform, etc, but here the motivation is different. The function a forges an association between data T, parameter θ, and auxiliary variable U. The point is that, once $T = t$ is observed, the very best possible inference about θ is obtained if and only if the corresponding U value is observed. As U is, by construction, unobservable, the inference problem can be recast into one of accurately guessing or predicting the unobserved U. This is where random sets will come in handy.

3 Random sets

3.1 A general overview

Let $\mathbb{U}$ be a space and $\mathcal{S}$ a random set, taking values in a collection of subsets of $\mathbb{U}$, with distribution $\mathsf{P}_\mathcal{S}$. There is an rigorous theory for random sets, presented beautifully in,

e.g., Molchanov (2005). Our case here turns out to be a relatively simple special case of the general theory so, for the sake of simplicity, we shall ignore the various topological and measure-theoretical technicalities that appear in more formal treatments.

There are a variety of related ways to describe the distribution of the random set. One approach is via the capacity functional (Molchanov 2005, Def. 1.4). Another description, popular in applications involving uncertainty (e.g., statistics, artificial intelligence, etc), is the belief function, $\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\}$, $K \subseteq \mathbb{U}$. One can easily see that the belief function is a formal analogue to the distribution function of a random variable. One key difference when dealing with random sets, compared to random variables, is that the complementation law generally fails, i.e., $\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} + \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K^c\} \leq 1$, with equality for all K if and only if $\mathcal{S}$ is a singleton set with $\mathsf{P}_{\mathcal{S}}$-probability 1. One can discuss belief functions without explicitly talking about random sets (e.g., Shafer 1979), though we shall not do so here. One particularly natural way that belief functions can emerge is through a sort of push-forward operation on a probability measure via a set-valued mapping; see, e.g., Dempster (1967), Nguyen (1978), and the discussion following the proof of Proposition 1 below. There is now a wide variety theoretical developments and applications of belief functions; see the volume edited by Yager and Liu (2008). In the remainder of this section, we shall focus only on those details that will be important in the sequel.

Consider now the special case where the random set is nested. In other words, the collection $\mathbb{S} \subset 2^{\mathbb{U}}$ of possible realizations of $\mathcal{S}$, called the support of $\mathcal{S}$, satisfies:

$$\text{for any } S, S' \in \mathbb{S}, \text{ either } S \subseteq S' \text{ or } S \supseteq S'. \tag{3}$$

In this case, the belief function corresponding to $\mathcal{S}$ is called consonant (Aregui and Denœux 2008; Balch 2012; Shafer 1976, 1987). Moreover, we may also conclude, via the usual continuity properties of the probability $\mathsf{P}_{\mathcal{S}}$, that the belief function is condensable. These together imply that the belief function (for $\mathcal{S}$) is fully characterized (see Shafer 1987, Sec. V.G) by the contour function, given by

$$f_{\mathcal{S}}(u) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \ni u\}, \quad u \in \mathbb{U}, \tag{4}$$

i.e., the probability that the random set $\mathcal{S}$ catches the fixed point $u \in \mathbb{U}$. As this is an ordinary function, not a set function, it will be easier to work with than the belief function. That this ordinary function captures the entire belief function can be seen by the formula

$$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = 1 - \sup_{u \in K^c} f_{\mathcal{S}}(u).$$

As we shall see in the next subsection, nested random sets, together with their corresponding contour functions (4) play an important role in this new theory.

3.2 Predictive random sets

In their investigations into the use of Dempster–Shafer theory for statistical inference, Martin et al. (2010) and Zhang and Liu (2011) observe that the corresponding belief functions have proper calibration properties only for certain classes of assertions or hypotheses. To rectify this mis-calibration, the previous authors argue that the Dempster–Shafer focal elements need to be enlarged, and that this can be accomplished by using

what are called predictive random sets. This combination of predictive random sets with the Dempster–Shafer theory of belief functions provides the mathematical backbone of a new approach, the so-called inferential model (IM) approach; for the complete details, see Martin and Liu (2013a). Here and in Section 4, we shall review this general theory with an emphasis on the construction of exact confidence regions.

Given the auxiliary space $\mathbb{U}$, equipped with measure P_U, let $\mathbb{S}$ be a collection of P_U-measurable subsets of $\mathbb{U}$. The collection $\mathbb{S}$ will serve as the support for the predictive random set; without loss of generality, we shall assume that $\mathbb{S}$ contains both $\varnothing$ and $\mathbb{U}$. Write $\mathcal{S}$ for the predictive random set, $\mathsf{P}_{\mathcal{S}}$ for its distribution, and $f_{\mathcal{S}}(u)$ for the corresponding contour function (4). An apparently new concept in the random set theory is that of validity. That is, the predictive random set $\mathcal{S}$ is *valid* if $f_{\mathcal{S}}(U)$ is stochastically no smaller than $\mathsf{Unif}(0,1)$ when $U \sim \mathsf{P}_U$. It will be shown in Section 4 that validity of the predictive random set leads to confidence regions with exact coverage probabilities.

Here, the interesting question is how to construct a predictive random set that satisfies this validity criterion. The answer is surprisingly simple. First, take $\mathcal{S}$ to be nested, so that its support $\mathbb{S}$ satisfies (3) and the belief function is consonant. As discussed in the previous subsection, this implies that the contour function $f_{\mathcal{S}}$ fully characterizes the distribution $\mathsf{P}_{\mathcal{S}}$. Now, since validity implicitly requires some connection between $\mathsf{P}_{\mathcal{S}}$ and P_U, our second condition should forge this connection. Indeed, we shall consider $\mathcal{S}$ with contour functions $f_{\mathcal{S}}$ that satisfy

$$f_{\mathcal{S}}(u) = \inf_{S \in \mathbb{S}: S \ni u} \mathsf{P}_U(S), \quad u \in \mathbb{U}. \tag{5}$$

We can now prove that these two conditions are sufficient for validity.

Proposition 1. *If $\mathcal{S}$ is nested and its contour function satisfies* (5)*, then it is valid.*

Proof. Pick any $\alpha \in (0,1)$ and set $S_\alpha = \bigcup\{S \in \mathbb{S} : \mathsf{P}_U(S) \leq \alpha\}$, the largest $S \in \mathbb{S}$ with P_U-probability no more than α. Based on (5), we can easily see that $f_{\mathcal{S}}(u) \leq \alpha$ if and only if $u \in S_\alpha$. Therefore, $\mathsf{P}_U\{f_{\mathcal{S}}(U) \leq \alpha\} = \mathsf{P}_U(S_\alpha) \leq \alpha$. This holds for all α, so $f_{\mathcal{S}}(U)$ is stochastically no smaller than $\mathsf{Unif}(0,1)$, and the claimed validity holds. ☐

Nested predictive random sets are simple to construct. For example, suppose P_U is a $\mathsf{Unif}(0,1)$ distribution and define a predictive random set $\mathcal{S}$ given by

$$\mathcal{S} = \{u : |u - 0.5| \leq |U - 0.5|\}, \quad \text{with} \quad U \sim \mathsf{P}_U. \tag{6}$$

Then the support $\mathbb{S}$ of $\mathcal{S}$ contains all symmetric intervals S centered at 0.5 of width less than or equal to 1, which is clearly a nested collection. Next, consider the distribution $\mathsf{P}_{\mathcal{S}}$ inherited from P_U, i.e.,

$$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = \mathsf{P}_U\bigl\{\{u : |u - 0.5| \leq |U - 0.5|\} \subseteq K\bigr\}.$$

With a little effort, the reader can easily convince his/herself that $\mathsf{P}_{\mathcal{S}}$ satisfies (5). Therefore, $\mathcal{S}$ satisfies the conditions of Proposition 1 and, hence, is valid. In fact, validity of $\mathcal{S}$ can be shown directly by checking that the contour function $f_{\mathcal{S}}(\cdot)$ in (4) satisfies $f_{\mathcal{S}}(U) \sim \mathsf{Unif}(0,1)$ for $U \sim \mathsf{P}_U$. Martin and Liu (2013a) refer to (6) as the "default" predictive random set.

The previous arguments can be generalized. For example, let P_U be a general non-atomic distribution on $\mathbb{U}$ and take h to be a continuous nowhere constant function from $\mathbb{U}$ to $\mathbb{R}$. Then it follows similarly that the predictive random set $\mathcal{S}$, defined by

$$\mathcal{S} = \{u : h(u) \leq h(U)\}, \quad \text{with} \quad U \sim \mathsf{P}_U, \tag{7}$$

is admissible and, hence, valid. In many scalar parameter problems, by performing suitable auxiliary variable transformations, one can get $\mathbb{U} = (0, 1)$ and $\mathsf{P}_U = \mathsf{Unif}(0, 1)$, so that the default predictive random set (6) can be used. However, this more general construction of a valid predictive random set proves useful in cases where the auxiliary variable space $\mathbb{U}$ is of dimension two or more.

4 Plausibility regions

4.1 Plausibility functions for statistical inference

Recall the auxiliary variable representation of the sampling model, i.e., $T = a(U, \theta)$, where T is the statistic of interest, and $U \sim \mathsf{P}_U$. Let $T = t$ be the observed statistic. If the auxiliary variable U were also observed, say $U = u$, then the best possible inference on θ could be obtained, and would be represented by the set

$$\Theta_t(u) = \{\theta : t = a(u, \theta)\}.$$

This set could be a singleton, but need not be. The idea is that *if* the auxiliary variable were observed, then given $T = t$, one can solve for the parameter of interest, and $\Theta_t(u)$ is exactly this set of solutions. In other words, $\Theta_t(u)$ defines a t-dependent compatibility relation (Shafer 1987) on Θ and $\mathbb{U}$.

Since the auxiliary variable U is not observable, it is not clear exactly how we should make use of the sets $\Theta_t(u)$. In the classical Dempster–Shafer context, a belief function is defined on Θ by pushing the measure P_U on $\mathbb{U}$ forward through the t-dependent set-valued mapping $\Theta_t(\cdot)$, creating a new random set $\Theta_t(U)$, with $U \sim \mathsf{P}_U$. But as we indicated above in Section 3.2, the Dempster–Shafer belief functions, generally, are not properly calibrated, and here is where the predictive random set $\mathcal{S}$ comes into play. The validity property for $\mathcal{S}$ ensures that it will hit its target—a draw from P_U—with large $\mathsf{P}_{\mathcal{S}}$-probability. Therefore, we may push the measure $\mathsf{P}_{\mathcal{S}}$, or its corresponding belief function forward, via the map $\Theta_t(\cdot)$, to obtain the bigger random set

$$\Theta_t(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_t(u) \tag{8}$$

The intuition is that we expect $\Theta_t(\mathcal{S})$ to contain the true θ with large $\mathsf{P}_{\mathcal{S}}$-probability. So, we understand $\Theta_t(\mathcal{S})$ as a (random) set of "reasonable" guesses of θ: for a given $A \subseteq \Theta$, if $\Theta_t(\mathcal{S}) \cap A \neq \varnothing$, then we cannot rule out the possibility that the true θ resides in A. By the *plausibility* of A we mean the $\mathsf{P}_{\mathcal{S}}$-probability that $\Theta_t(\mathcal{S}) \cap A \neq \varnothing$,

$$\mathsf{pl}_t(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_t(\mathcal{S}) \cap A \neq \varnothing\}, \quad A \subseteq \Theta. \tag{9}$$

We shall refer to $\mathsf{pl}_t(A)$ as the plausibility function at A; though the notation does not reflect this, the reader should keep in mind that pl_t depends on $\mathsf{P}_{\mathcal{S}}$.

A few remarks about the above construction are in order. First, we could have equivalently started by defining a belief function $\mathsf{bel}_t(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_t(\mathcal{S}) \subseteq A\}$ for the random set $\Theta_t(\mathcal{S})$ and then the plausibility function $\mathsf{pl}_t(A) = 1 - \mathsf{bel}_t(A^c)$. We will not need the belief function in what follows, so the presentation here is more direct. Second, the fact that the new random set in (8) is generally larger than that of the classical Dempster–Shafer analysis leads to smaller belief functions. It is this squashing of the Dempster–Shafer belief function or, equivalently, the boosting of the plausibility function, that accounts for the improved calibration. Indeed, as we show below, if the predictive random set is valid, then the squashing/boosting will be just enough to attain the desired calibration. Third, the argument here for combining $\Theta_t(\cdot)$ with $\mathcal{S}$ as in (8) is just a special case of Dempster's rule of combination (Dempster 1967, 2008), though writing out the details formally perhaps does not provide any additional insight. The key point is that Dempster's argument does not require that uncertainty on the $\mathbb{U}$-space be summarized with a genuine probability measure. In particular, the same line of reasoning applies if uncertainty on $\mathbb{U}$ is described via a belief function, like in our present case.

If $\mathsf{P}_{\mathcal{S}}\{\Theta_t(\mathcal{S}) = \varnothing\} > 0$, then one must adjust the formula (9) by conditioning on the event that $\Theta_t(\mathcal{S}) \neq \varnothing$. To avoid such conditioning here, we assume that

$$\Theta_t(u) \neq \varnothing \quad \text{for all } (t, u) \text{ pairs.} \tag{10}$$

This assumption essentially boils down to there being no non-trivial constraints on the parameter θ in the sampling model $\mathsf{P}_{T|\theta}$. An example of a non-trivial constraint is in a Poisson problem where the mean θ has a positive lower bound. Most regular problems, including the examples in Section 5, satisfy (10), though there are some that do not. Assumption (10) is not necessary to construct plausibility regions with the desired coverage probabilities, but it will make our presentation easier. The correction requires a relatively technical kind of stretching of the predictive random sets to maintain validity; see Ermini Leaf and Liu (2012).

For the important special case where $A = \{\theta\}$ is a singleton, we write $\mathsf{pl}_t(\theta) = \mathsf{pl}_t(\{\theta\}) = \mathsf{P}_{\mathcal{S}}\{\Theta_t(\mathcal{S}) \ni \theta\}$. Note that this special plausibility function is just the contour function (4) corresponding to the new nested random set $\Theta_t(\mathcal{S})$. This plausibility function also gives rise to the $100(1-\alpha)\%$ *plausibility region*:

$$\mathscr{P}_\alpha(t) = \{\theta : \mathsf{pl}_t(\theta) > \alpha\}. \tag{11}$$

As we demonstrate below, if $\mathcal{S}$ is valid, then the plausibility region $\mathscr{P}_\alpha(T)$ has coverage probability at least $1-\alpha$ and, in many cases, equality is attained.

4.2 Coverage probability results

The first result gives shows that $\mathsf{pl}_T(\theta)$, for $T \sim \mathsf{P}_{T|\theta}$, is stochastically no smaller than $\mathsf{Unif}(0,1)$ under mild conditions. From this, plausibility region's advertised attainment of the nominal coverage coverage probability follows easily.

Proposition 2. *Fix $\theta \in \Theta$. If $\mathcal{S}$ is valid and* (10) *holds, then $\mathsf{pl}_T(\theta)$ is stochastically no smaller than* $\mathsf{Unif}(0,1)$ *when $T \sim \mathsf{P}_{T|\theta}$.*

Proof. From the alternative description of the sampling model $\mathsf{P}_{T|\theta}$ in (2), for $T \sim \mathsf{P}_{T|\theta}$, there exists a corresponding $U_T \sim \mathsf{P}_U$ such that $T = a(\theta, U_T)$. Moreover, it follows easily from the definition of $\Theta_t(\mathcal{S})$ that $\Theta_T(\mathcal{S}) \ni \theta$ if and only if $\mathcal{S} \ni U_T$. Therefore, $\mathsf{pl}_T(\theta) = \mathsf{P}_{\mathcal{S}}\{\Theta_T(\mathcal{S}) \ni \theta\} = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \ni U_T\} = f_{\mathcal{S}}(U_T)$. Since $\mathcal{S}$ is valid, $f_{\mathcal{S}}(U_T)$ is stochastically no smaller than $\mathsf{Unif}(0,1)$, as a function of $U_T \sim \mathsf{P}_U$, and so the claim follows. $\square$

There are two relevant results that can be derived from Proposition 2 and its proof.

- The first is that, for any $\alpha \in (0,1)$, the coverage probability of the plausibility region $\mathscr{P}_\alpha(T)$ in (11) is at least $1-\alpha$, i.e., $\mathsf{P}_{T|\theta}\{\mathscr{P}_\alpha(T) \ni \theta\} \geq 1-\alpha$ for all θ. To see this, note that $\mathsf{P}_{T|\theta}\{\mathscr{P}_\alpha(T) \ni \theta\} = \mathsf{P}_{T|\theta}\{\mathsf{pl}_T(\theta) > \alpha\}$. By Proposition 2, $\mathsf{pl}_T(\theta)$ is stochastically no smaller than $\mathsf{Unif}(0,1)$. This implies that the latter probability is no smaller than $\mathsf{P}\{\mathsf{Unif}(0,1) > \alpha\} = 1-\alpha$, hence the claim.
- Second, confidence regions can be constructed directly from $\Theta_t(\cdot)$ and the support sets $S \in \mathbb{S}$ for the predictive random sets. Indeed, for fixed $S \in \mathbb{S}$, we know from the above proof that $\Theta_T(S) \ni \theta$ if and only if $U_T \in S$. So, $\mathsf{P}_{T|\theta}\{\Theta_T(S) \ni \theta\} = \mathsf{P}_U(S)$, and if we select S such that $\mathsf{P}_U(S) = 1-\alpha$, then $\Theta_t(S)$ is a $100(1-\alpha)\%$ confidence region for θ. Therefore, an alternative $100(1-\alpha)\%$ confidence region construction selects the smallest S with $\mathsf{P}_U(S) = 1-\alpha$ and takes $\mathscr{C}_\alpha(t) = \Theta_t(S)$.

An important question is, under what conditions, is the coverage probability exactly equal to $1-\alpha$ or, equivalently, when is $\mathsf{pl}_T(\theta)$, with $T \sim \mathsf{P}_{T|\theta}$, exactly uniformly distributed? It turns out that there are two conditions needed. First, T must have a continuous distribution $\mathsf{P}_{T|\theta}$, otherwise $\mathsf{pl}_T(\theta)$ cannot be continuous. Second, the predictive random set must be *exact*, not just valid, i.e., $f_{\mathcal{S}}(U)$ must be exactly $\mathsf{Unif}(0,1)$ for $U \sim \mathsf{P}_U$. This exactness property holds for the default predictive random set (6) and its generalized version (7). Therefore, for problems with continuous T, if we choose an exact predictive random set, such as one of those in (6) or (7), then the plausibility region $\mathscr{P}_\alpha(T)$ has coverage probability exactly $1-\alpha$.

5 Examples

5.1 Power law process

Consider a continuous time non-homogenous Poisson process $\{N_y : y \geq 0\}$, where the mean function $m(y) = \mathsf{E}(N_y)$ satisfies $m(y) = \psi y^\theta$, for $\psi, \theta > 0$. Such a process is called a power law process (e.g., Gaudoin et al. 2006). The parameter ψ is a scale parameter and θ is a shape parameter. Though both ψ and θ are unknown, the goal here is to construct a plausibility interval for θ based on n observed event times $Y_1 \leq \cdots \leq Y_n$.

For these data, the log-likelihood function for (ψ, θ) looks like

$$\ell(\psi,\theta) = n \log \psi + n \log \theta + (\theta - 1) \sum_{i=1}^n \log Y_i - \psi Y_n^\theta.$$

By the Neyman–Fisher factorization theorem, a joint sufficient statistic for (ψ, θ) is the pair $(\sum_{i=1}^n \log Y_i, Y_n)$. This sufficient statistic is a one-to-one transformation of the max-

imum likelihood estimator $(\hat{\psi}, \hat{\theta})$, given by

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n-1} \log(Y_n/Y_i)} \quad \text{and} \quad \hat{\psi} = n/Y_n^{\hat{\theta}}.$$

Therefore, $(\hat{\psi}, \hat{\theta})$ is a minimal sufficient statistic for (ψ, θ). Moreover, for all ψ, the vector $\{\log(Y_n/Y_{n-1}), \ldots, \log(Y_n/Y_1)\}$ is distributed as a sorted sample of $n-1$ independent random variables from an exponential distribution with mean $1/\theta$ (e.g., Crow 1974). For simplicity, we take

$$T = n/\hat{\theta} = \sum_{i=1}^{n-1} \log(Y_n/Y_i),$$

which has a gamma distribution with shape $n-1$ and scale $1/\theta$. Some information about θ is lost by ignoring the $\hat{\psi}$ component of the joint minimal sufficient statistic, but the required marginalization strategy is beyond our present scope; see Martin and Liu (2013b). So, we shall consider here the simple association

$$T = F_{n-1,1/\theta}^{-1}(U), \quad U \sim \mathsf{Unif}(0,1),$$

where $F_{n-1,1/\theta}$ denotes the gamma distribution function with shape $n-1$ and scale $1/\theta$. If the default predictive random set $\mathcal{S}$ in (6) is used for U, then the plausibility function turns out to be

$$\mathsf{pl}_t(\theta) = 1 - |2F_{n-1,1/\theta}(t) - 1|, \quad \theta > 0,$$

which can be readily evaluated numerically. Then the $100(1-\alpha)\%$ plausibility interval $\mathscr{P}_\alpha(t)$ for θ is given by

$$\mathscr{P}_\alpha(t) = \{\theta : \mathsf{pl}_t(\theta) > \alpha\} = \{\theta : \alpha/2 < F_{n-1,1/\theta}(t) < 1 - \alpha/2\}.$$

Since $1/\theta$ is a scale parameter in $F_{n-1,1/\theta}(t)$, the right-hand side above can be rewritten as $\{\theta : \alpha/2 < F_{n-1,1}(\theta t) < 1-\alpha/2\}$. Therefore, if we let $\gamma_{n-1,1}(q)$ denote the qth quantile of the gamma distribution with shape $n-1$ and scale 1, then the plausibility interval can be written as a genuine interval,

$$\mathscr{P}_\alpha(t) = \Bigl(\frac{\gamma_{n-1,1}(\frac{\alpha}{2})}{t}, \frac{\gamma_{n-1,1}(1-\frac{\alpha}{2})}{t}\Bigr).$$

This is equivalent to the exact confidence interval given in equation (6) of Gaudoin et al. (2006) in terms of chi-square quantiles.

5.2 Exponential regression through the origin

Consider a special case of an exponential log-linear model, where $Y_1, \ldots, Y_n$ are independent exponential random variables and Y_i has mean $e^{\theta x_i}$, $i = 1, \ldots, n$, for fixed covariates $x_1, \ldots, x_n$. The goal is to produce a plausibility interval for the slope parameter θ.

The log-likelihood function for θ looks like

$$\ell(\theta) = -\sum_{i=1}^{n} (\theta x_i + e^{\log Y_i - \theta x_i}),$$

and the likelihood equation is given $\sum_{i=1}^n (e^{\log Y_i - \theta x_i} - 1)x_i = 0$. Let T be the solution to this equation, the maximum likelihood estimator of θ. If $G_\theta = G_{\theta,x,n}$ is the distribution function of T, then a suitable association is, again, given by

$$T = G_\theta^{-1}(U), \quad U \sim \mathsf{Unif}(0,1).$$

The distribution function G_θ is not available in closed form, but it can be evaluated via Monte Carlo. If the default predictive random set $\mathcal{S}$ in (6) is used for U, then again the plausibility function for θ is of the form

$$\mathsf{pl}_t(\theta) = 1 - |2G_\theta(t) - 1|, \quad \theta \in \mathbb{R}.$$

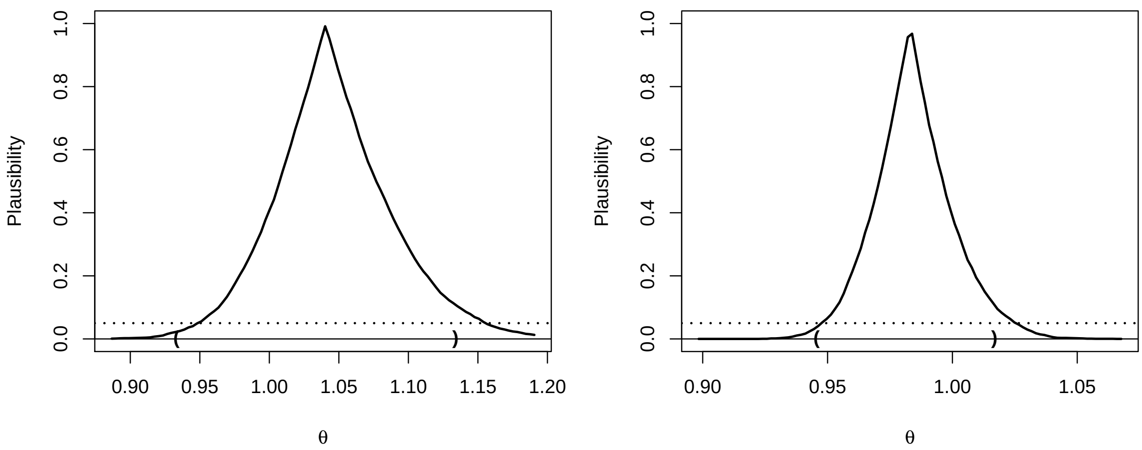
No expressions are available for the plausibility function in this case, but, again, it is relatively easy to evaluate numerically via Monte Carlo.

For illustration, Figure 1 displays plots of the plausibility function $\mathsf{pl}_t(\theta)$, as a function of θ, for two simulated data sets, one of size $n = 10$, the other of size $n = 20$. Here the covariate values are $x_i = i$, $i = 1, \ldots, n$, and the true parameter value is $\theta = 1$. This function is evaluated by a Monte Carlo integration step performed at each point θ on the horizontal axis. For comparison, the endpoints of the 95% confidence interval based on asymptotic normality of the maximum likelihood estimator are also displayed. In both cases, the two intervals are comparable, which is to be expected. However, for such small n, it is unlikely that the asymptotic normality has kicked in, so the actual coverage probability of the latter is likely different from the target 0.95. The plausibility interval, on the other hand, has coverage probability exactly equal to 0.95 based on the theory developed in Section 4.2. Indeed, in a simulation of 5000 data sets of size $n = 10$, under same setup as above, the estimated coverage probabilities for the exact plausibility interval and asymptotic confidence interval are 0.951 and 0.934, respectively.

6 Discussion

In this paper, we discuss a new approach for the construction of confidence regions based on the theory of random sets. The key result is that if the predictive random set $\mathcal{S}$ for the unobservable auxiliary variable U is valid, in the sense that it misses its its target not too often, then the corresponding plausibility region has at least the nominal coverage probability. It is important that this validity result is not asymptotic and, moreover, does not depend on any characteristic of the problem that is unknown. Therefore, it is generally quite easy to specify a valid predictive random set, and a default choice is given here and used in several examples to obtain practically useful results.

Here the focus was on simplicity rather than generality. Though the two examples involved only scalar parameter, essentially the same strategy would apply for a multi-parameter problem. A challenging problem in multi-parameter situations is to give an exact confidence region for some component or, more generally, some scalar-valued function of the full parameter. This was the actual setup in the power-law process example in Section 5.1, though we sidestepped the main difficulty by ignoring a part of the minimal sufficient statistic. To incorporate all the information in the minimal sufficient statistic requires some careful manipulations which were beyond the present scope. A new and detailed look at such problems is given Martin and Liu (2013b).



(a) $n = 10$ (b) $n = 20$

Figure 1: Plausibility functions $\mathsf{pl}_t(\theta; \mathcal{S})$ versus θ in Section 5.2. Parentheses on the θ-axis mark the endpoints of the 95% confidence interval based on asymptotic normality of the maximum likelihood estimator. Horizontal line at $\alpha = 0.05$ determines the endpoints of the 95% plausibility interval.

The primary goal here was to construct confidence regions that attain the nominal coverage probability. We found that, in many cases, including the two examples in Section 5, the plausibility regions will actually hit this target on the nose. A natural follow-up question is if these plausibility regions are "optimal" in some sense, i.e., do the plausibility regions have smallest average size, say, among all those regions that hit the desired coverage probability? This question is the focus of ongoing investigations.

Acknowledgments

The author is thankful for comments and suggestions given by Chuanhai Liu. This work is partially supported by the U.S. National Science Foundation, DMS–1208833.

References

Aregui, A. and Denœux, T. (2008), "Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities," *Internat. J. Approx. Reason.*, 49, 575–594.

Balch, M. S. (2012), "Mathematical foundations for a theory of confidence structures," *Internat. J. Approx. Reason.*, 53, 1003–1019.

Crow, H. L. (1974), "Reliability analysis for complex, repairable systems," in *Reliability and Biometry*, eds. F., P. and Serfling, R. J., Philadelphia: Society for Industrial and Applied Mathematics (SIAM), pp. 379–410.

Dempster, A. P. (1967), "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, 38, 325–339.

— (2008), "Dempster–Shafer calculus for statisticians," *Internat. J. of Approx. Reason.*, 48, 265–277.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Ermini Leaf, D. and Liu, C. (2012), "Inference about constrained parameters using the elastic belief method," *Internat. J. Approx. Reason.*, 53, 709–727.

Gaudoin, O., Yang, B., and Xie, M. (2006), "Confidence intervals for the scale parameter of the power-law process," *Comm. Statist. Theory Methods*, 35, 1525–1538.

Martin, R. and Liu, C. (2012), "Conditional inferential models: combining information for prior-free probabilistic inference," Unpublished manuscript, `arXiv:1211.1530`.

— (2013a), "Inferential models: A framework for prior-free posterior probabilistic inference," *J. Amer. Statist. Assoc.*, to appear, `arXiv:1206.4091`.

— (2013b), "Marginal inferential models: optimal prior-free probabilistic inference on interest parameters," Unpublished manuscript.

Martin, R., Zhang, J., and Liu, C. (2010), "Dempster–Shafer theory and statistical inference with weak beliefs," *Statist. Sci.*, 25, 72–87.

Molchanov, I. (2005), *Theory of random sets*, Probability and its Applications (New York), London: Springer-Verlag London Ltd.

Nguyen, H. T. (1978), "On random sets and belief functions," *J. Math. Anal. Appl.*, 65, 531–542.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton, N.J.: Princeton University Press.

— (1979), "Allocations of probability," *Ann. Probab.*, 7, 827–839.

— (1987), "Belief functions and possibility measures," in *The Analysis of Fuzzy Information, Vol. 1: Mathematics and Logic*, ed. Bezdek, J. C., CRC, pp. 51–84.

Yager, R. and Liu, L. (eds.) (2008), *Classic works of the Dempster-Shafer theory of belief functions*, vol. 219, Berlin: Springer.

Zhang, J. and Liu, C. (2011), "Dempster–Shafer inference with weak beliefs," *Statist. Sinica*, 21, 475–494.